SmaCS: Smart Classification System for the Design, Maintenance and Use of Complex Terminologies. Application in Pediatric Cardiology.

Freek van den Heuvel MD[1, 2], Teun Timmers Ph.D.[2], John Hess MD Ph.D.[1]
[1]Division of Pediatric Cardiology, Sophia Children's Hospital Rotterdam
[2]Department of Medical Informatics, Erasmus University Rotterdam
The Netherlands

The development of a generic Smart Classification System (SmaCS) for the design, construction, maintenance, and use of complex controlled terminologies is described. The ability of SmaCS to create and maintain terminologies that are combinations of structured terms and domain-specific knowledge is an important aspect of its design. SmaCS can therefore be used both for both controlled data-collection, using integrity rules included in the terminology, and intelligent data-retrieval. The design and implementation of SmaCS and application in the domain of pediatric cardiology are described.

INTRODUCTION

Over the past two decades, developments in the management of patients with *congenital heart disease* (CongHD) have substantially improved both short- and medium-term results. A relative lack of insight into the long-term effects of CongHD treatment in general and the influence of a continuous introduction of new and improved techniques for the management of specific abnormalities stress the need for clinical evaluation studies and development of management protocols. To support such follow-up research, a domain specific database management system was developed, utilizing a vocabulary of CongHD terminology for the coding of often complex morphologic abnormalities and follow-up events in individual patients.[1] The system has been used by the Dutch centers for Pediatric Cardiology since 1988. However, attempts to explore the data that had been collected since its introduction proved to be problematic. This was mainly due to the complexity of the CongHD domain, implemented as an extensive but relatively unstructured and static vocabulary of terms, and inherent limitations of the system. The impossibility to extract research populations from a large collection of patient data with complex diagnostic and therapeutic codings and lacking facilities for maintenance and use of the vocabulary for data-retrieval were identified as the most prominent restrictions. Based on these experiences we have developed a prototype Smart Classification System, SmaCS, to provide a tool for the creation, maintenance, and use of complex terminologies. The considerations underlying its design, their implementation, and application of SmaCS in follow-up research in the domain of CongHD are described in this paper.

DESIGN CONSIDERATIONS

The considerations for the design of SmaCS were based both on our experiences with the above-mentioned system and the design criteria for controlled medical terminologies as described by Cimino et al.[2,3] Our most important design criteria are:

- Domain independence. A generic system for the construction of (medical) terminologies irrespective of their application domain and complexity, required an approach where the structures underlying specific terminologies could be specified separately and independently from each other. In such a system concepts with their aspects and relations between concepts are used to construct a concept model for the representation of the structure and characteristics of a specific terminology.
- Integration of domain specific knowledge into the terminology: Two mechanisms were introduced to accomplish this: Firstly, specification of inter-term relationships to relate the individual items a terminology consists of in a logical and medically correct way; Secondly, specification of *integrity rules* (IRs) as aspects of individual terms in order to enable integrity checking during data-entry and to facilitate retrieval.
- Representation of terms from a single terminology into different views. In this way organization of terms from the same terminology can be optimized for specific purposes. This is necessary e.g. when a terminology organized according to pathology is used for data-collection where representation of terms according to clinical practice rather than pathology is required. Within each view, arrangement of terms and specification of properties of individual terms must be possible independent from the organization of the underlying terminology.
- Tools for maintenance and extension of terminologies and their views. This should include mechanisms for the management of obsolete terms as these must be preserved to ensure the integrity of

0195-4210/95/$5.00

historic data. Application of these terms should be limited to the data-retrieval process.
- Graphical presentation to facilitate representation of large terminologies with complex and multiple relationships between individual terms.

In addition to this SmaCS should incorporate a module to support the analysis of patient data with extensive codings. This module should enable:
- Stepwise approach to the solution of complex research questions: Each analysis step represents a sub-population of patients from a patient database based on either a collection of terms (term-set) gathered using SmaCS or a list of patients (patient-set). Each analysis step must be accessible for inspection to have insight into and control over the result.
- Storage and retrieval of queries. In this way previously defined queries can be recalled for further processing or to re-run a research question at a different moment or on a different database.

DESIGN & IMPLEMENTATION

Technical design, classification system

To meet the criteria described in the previous section, a generic classification structure was designed consisting of three layers that together form the SmaCS terminology database. These layers are the Conceptual Classification Schema (CCS) providing a generic mechanism for specification of structures underlying terminologies; the Classification Layer (CL); and the view layer (VL). Both the CL and VL offer mechanisms (relationships between terms and integrity rules) for the inclusion of domain knowledge. The VL also enables restructuring of terms into different organizations suit specific purposes (e.g. presentation according to clinical entities encountered in clinical practice when a terminology is used for data-entry, or according to surgical procedures when data-retrieval is performed). The individual layers with their characteristics are described in the following sections. All objects in the layers are stored in a relational database management system.

The Conceptual Classification Schema. A CCS is a directed graph, providing the meta schema on which terminologies and their views are based. In this layer a semantic network model is built from the various nodetypes and linktypes that are formal representations of the individual terms (in this context called nodes) and their relationships (links), in a given terminology. Characteristics (in this context called attributes) common to all nodes, such as its name, identifier or synonyms, are implemented as aspects common to all nodetypes. Among them is a special type of aspect: the User Defined Attribute (UDA). This aspect allows for definition of attributes that are characteristic of and limited to the nodes from a specific terminology. Similarly, linktypes are formal representations of the links present between the nodes in a terminology. Linktypes can be specified between different nodetypes but also to represent links between nodes instantiated from the same nodetype. Such a linktype constitutes a cycle. In Figure 1 the definitions and structure of database tables used for the storage of the various CCS objects are shown. In a similar way the various objects of the classification- and the view-layer are implemented.

The Classification Layer. Each classification is based on a single CCS. Several classifications can be built using the same CCS. Nodes can only be instantiated from the nodetypes defined in the CCS on which the classification is built. Similarly, links can only be specified between nodes that are instantiated from the nodetypes specified in the linktype definition. For each node, its name, a (unique) identifier and the SELECTABLE attribute are obligatory. This last attribute is necessary to specify whether the node is selectable or non-selectable. In our congenital heart disease classification, both generalising nodes, i.e. nodes that have children, and nodes that represent obsolete terminology are stored as non-selectable. This attribute indicates whether the term may be selected during data-entry.

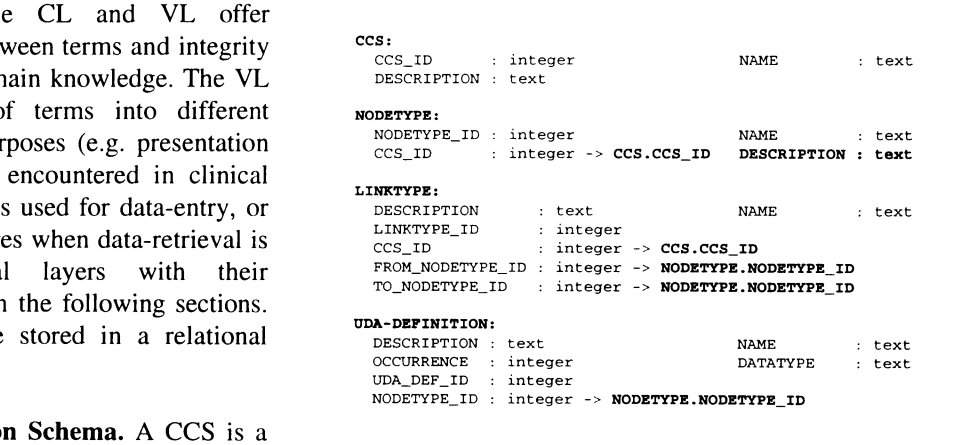

Figure 1: The main tables and their relations (right arrows) used for the storage of nodetypes and linktypes of CCS definitions. For each UDA its type (integer, text, or date) and behavior (single value/list and optional/obligatory) are described by the DATATYPE and OCCURRENCE fields respectively.

Another, optional, attribute can be used for the specification of an *integrity rule* (IR) associated with a specific node. The syntax for specification of IRs

consists of a collection of logical operators used in conjunction with node identifiers. Together with each IR a message is specified that is shown in case an IR is violated. In Figure 2 a small part of a classification consisting of terminology used in the domain of CongHD is shown to illustrate the various SmaCS components and their presentation to the user. In Figure 3 an example is shown of an IR that is specified at a (generalizing) node in the classification branches shown in Figure 2.

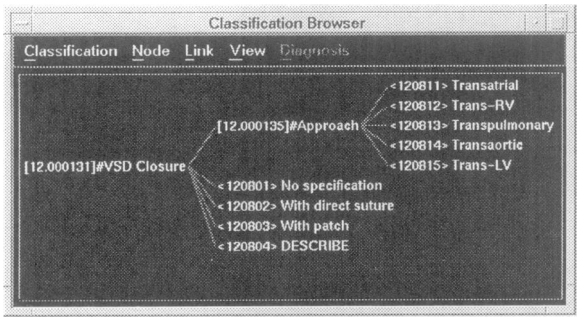

Figure 2: Part of the CongHD terminology. The items shown in this branch represent terms used for the description of surgical closure of a ventricular septal defect (VSD). Numbers in front of the descriptions represent the unique identifiers of the nodes. A # symbol indicates the presence of an integrity rule at that node. An example of the contents and function of a rule present in this branch is shown in Figure 3.

Rule at node 12.000131:
EXACT(1, 120801 120802 120803 120804) AND 12.000135

A valid selection consists of exactly one node selected from the terminal nodes that are the children of node 12.000131 in combination with a valid selection made at node 12.000135.

Figure 3: Example of an integrity rule specified at the node with identifier 12.000131 in Figure 2.

The View Layer. Views are alternative representations and extensions of a specific classification. Therefore views are based on classifications rather than on a CCS. All nodes within a view are part of the underlying classification. A view node inherits its characteristics from their classification counterparts, but these aspect values can be changed or extended. In many cases separate names and IRs are provided because view nodes may have children that are different from their corresponding classification nodes. Whereas view nodes are extensions of classification nodes, this is not true for view links. A view link is directly associated with a linktype in the CCS. An example illustrating the necessity for this is a link-type PART_OF_SYNDROME which is a link type defined in the CCS that is exclusively used in the various clinical views that have been defined on top of our CongHD terminology. This CongHD terminology is structured using morphological and pathological relationships between the various nodes, and no links for the representation of involvement of nodes in specific syndromes as they are encountered in clinical practice are present. An example of a view is shown in Figure 4

Technical design, query system

In order to achieve the required flexibility of data-retrieval, the SmaCS query system was designed to specify a specific research question as a logical combination of discrete small sub-questions. In this approach each query is made up of two sections:
The first section contains the various sub-questions contained in the research question to be solved. These sub-questions represent simple selections from the patient database based on either a term-set selected from a terminology using the classification system part of SmaCS or a patient-set. Such a patient-set may consist of a simple list of patient identifiers or contain the result of a query based on administrative or follow-up patient data. The results of each sub-question based on either a code- or patient-set can be inspected individually to provide feedback and monitor the final outcome at the various steps. The second section provides means for combination of the various sub-questions into an actual research question using logical operators. Individual query definitions as well as the results of a query can be stored for later or further processing. In Figure 5 an example of a research question is shown.

Implementation

The layers of the SmaCS terminology database (CCS-, Classification-, and View-Layer) are implemented in a data-manager layer. This data-manager layer is realized using the relational database management system INGRES. Both the tables and the relations specified among them together with the database rules necessary for the maintenance of the terminology database are physically stored in the SmaCS database. These database specific routines are realized using INGRES specific tools and are triggered whenever changes to the database contents are made by inserting, updating, and deleting of tuples. In this way the integrity of the database is always assured, irrespective of how the SmaCS terminology database is approached. This has the advantage that the SmaCS terminology database can be used separately from the other layers of the SmaCS prototype. This makes it

possible to use a classification as a component in a dedicated patient database management system separately from the SmaCS development system. On top of this data-manager layer an intermediary layer and user-interface layer were designed and implemented in C and MOTIF on an HP9000 UNIX workstation. The intermediary layer controls the communication and error-handling between the data-manager and (graphical) user-interface layer. The query module was implemented as a separate application that used the SmaCS as a 'terminology server'.

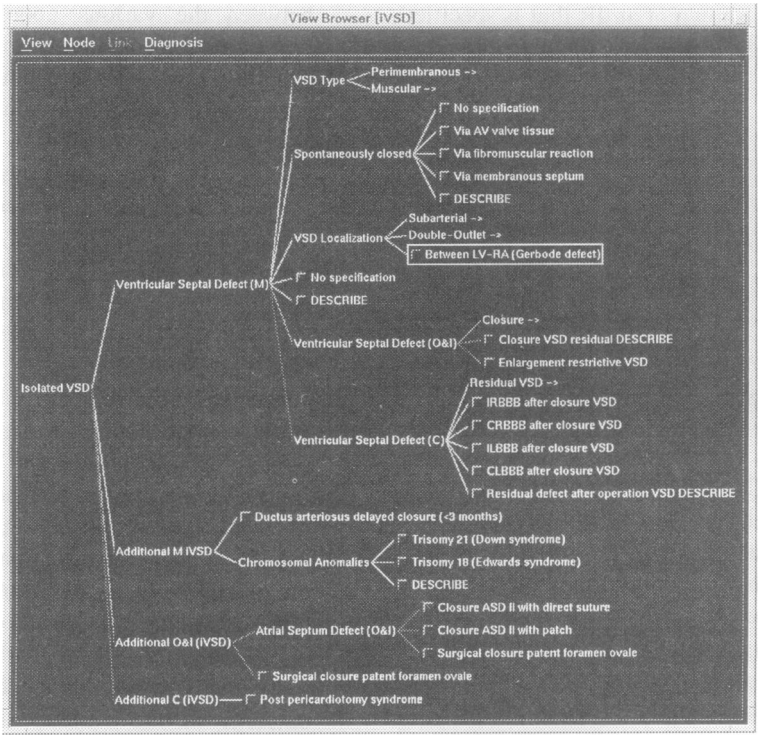

Figure 4: View representing the clinical entity 'Isolated VSD' (ventricular septal defect). In this view not only the morphological aspects of this frequently encountered abnormality are present, but also surgical procedures, complications and associated abnormalities that apply to this entity are represented.

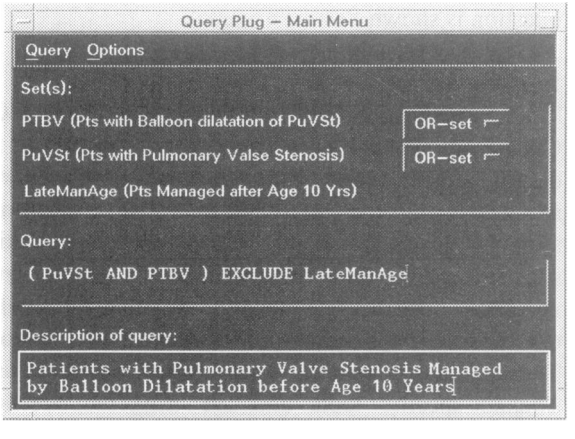

Figure 5: Construction of a research query using the SmaCS query module. In the upper part of the window two code-sets and a patient-set each representing a sub-selection from the patient database have been specified. In the middle part of the window these sets are combined into the final selection.

APPLICATION FOR CongHD TERMINOLOGY

Using the SmaCS system, an extensive controlled terminology was created, based on an existing vocabulary for the description of morphologic, diagnostic and treatment items in the domain of CongHD. This vocabulary, consisting of about 3800 relatively unstructured terms, was developed from the Brompton Hospital Code for the classification of CongHD[4] and has been used in a PC-based database system[1] in the Dutch centers for Pediatric Cardiology since 1988.

New terms to supplement missing items were added to the existing terms and structured into a new classification based on a relatively simple CCS. Using generalizing nodes and inter-term relationships related items were organised into branches together forming a directed acyclic graph. At several generalizing nodes, integrity rules were defined for the representation of medical knowledge with respect to co-occurrences and

exclusions among terms. On top of this new terminology, which was still organised according to morphologic principles, views were defined. Within these views, subsets of terms from the underlying terminology were reorganised to achieve representations of terminology into CongHD entities as they are encountered in clinical practice. For each view entity specific inter-term relations and IRs were specified. Currently the new terminology contains approximately 1750 terms, 4 views and 100 rules. Eventually, a terminology containing about 5000 terms, 45 views and 1000 rules is anticipated providing complete coverage of the CongHD domain. We estimate that such a complete implementation of the domain by a domain expert will take approximately 3 to 4 months.

Compared to other coding-systems, the CongHD terminology is a highly detailed coding system limited to a specialized domain in medicine. Overlap with existing systems is therefore limited. At more generic levels references to ICD and SNOMED codings are provided for exportation of data to registries and for administrative purposes. In the near future integration of SmaCS into a new CongHD follow-up information system is planned.

DISCUSSION

We expect that introduction of domain specific medical knowledge as represented by explicit relations between individual terms, specification of integrity rules, and organization of items from a terminology into entities that more closely represent clinical practice will improve completeness and correctness of data collected with this terminology.[5] In a similar way this knowledge substantially supports the retrieval of research populations from data that have been collected by grouping of relevant terms into database queries. The retrieval process is further enhanced by a stepwise approach in which research queries are solved by logical combinations of its elementary components.

Several studies on the principles and realization of systems for controlled terminologies have shown the importance and advantage of using abstract representations and inclusion of knowledge about terms in the creation, maintenance and use of such terminologies.[3,6,7,8,9] None of these however contain the notion of our integrity rules nor the property of reorganizing terminology into alternative representations (views). These features make SmaCS more suitable for the support of both data-collection and data-retrieval.

Currently SmaCS is used for the analysis of follow-up data of over 11.000 patients with CongHD collected in our center over a period of 7 years. Both the newly developed classification and query system have been successfully applied in a number of clinical studies concerning the management of patients with CongHD.[10]

References

1. Brower RW, Harinck E, Gittenberger-de Groot AC. A pediatric cardiology diagnostic coding system and database. In: Meester GT, Pinciroli F eds. Databases for cardiology. Dordrecht, the Netherlands: Kluwer Academic Publishers. 1991:259-71.
2. Cimino JJ, Hripcsak G, Johnson SB, Clayton PD. Designing an introspective, controlled medical vocabulary. In: Kingsland LW, ed. Proceedings of the 13th SCAMC. Washington DC: IEEE Computer Society Press. 1989:513-8.
3. Cimino JJ, Clayton PD, Hripcsak G, Johnson SB. Knowledge-based approaches to the maintenance of a large controlled medical terminology. J Am Med Informatics Assoc 1994;1:35-50.
4. The diagnosis of congenital heart diseases, incorporating the Brompton Hospital Code for the classification of congenital heart disease. Tunbridge Wells: Castle House Publications Ltd. 1985.
5. van den Heuvel F, Timmers T, van Mulligen EM, Hess J. Knowledge-based modelling for the classification and follow-up of patients with congenital heart disease. In: Lun KC, Degoulet P, Piemme TE, Rienhoff O eds. Proceedings of MEDINFO92. Amsterdam, the Netherlands: North-Holland. 1992:501-5.
6. Nowlan WA, Rector AL, Kay S, Horan B, Wilson A. A patient care workstation based on user centered design and a formal theory of medical terminology: PEN&PAD and the SKM formalism. In: Clayton PD ed. Proceedings of the 15th SCAMC. New York: McGraw-Hill. 1992:855-7.
7. Barr CE, Komorowski HJ, Pattison-Gordon E, Greenes RA. Conceptual modelling for the Unified Medical Language System. In: Greenes RA ed. Proceedings of the 12th SCAMC. Silver Spring MD: IEEE Computer Society Press. 1988:96-100.
8. Lindberg DAB, Humphreys BL, McCray AT. UMLS. The Unified Medical Language System. Meth Inform Med 1993;32:281-291.
9. Campbell KE, Das, AK, Musen MA. A logical foundation for representation of clinical data. J Am Med Informatics Assoc 1994:1:218:232.
10. van den Heuvel F, Timmers T, Hess J. Morphological, haemodynamic, and clinical variables as predictors for management of isolated ventricular septal defect. Br Heart J 1995;73:49-52.